图书情报工作
第 64 卷 第 5 期 2020 年 3 月


# 基于小数据的社交类学术 App 用户动态画像模型构建研究*

■ 张莉曼[1] 张向先[1] 吴雅威[1] 郭顺利[2]
[1] 吉林大学管理学院 长春 130022 [2]曲阜师范大学传媒学院 日照 276826

摘 要：[目的/意义]基于小数据构建社交类学术 App 用户动态画像模型，为社交类学术 App 平台有效预测用户行为演化趋势、提高精准服务水平提供思路和参考。[方法/过程] 首先，在深度剖析小数据概念及特点的基础上，结合社交类学术 App 特征，从用户表层行为和深层驱动因素两方面设计动态画像标签体系；其次，采集与用户强相关、高价值的小数据作为画像的数据支撑，并明确画像小数据的获取及处理方法；最后提出实现动态画像的研究方法并形成整体框架模型。[结果/结论] 基于小数据构建社交类学术 App 用户动态画像可有效细化画像粒度，改善以往画像滞后性弊端，对数据驱动情境下社交类学术 App 平台提升精准服务水平有重要的参考价值。
关键词：小数据 社交类学术 App 用户动态画像 行为预测
分类号：G250
DOI：10.13266/j.issn.0252－3116.2020.05.006



随着互联网的发展和信息的快速更新，传统的线下交流模式已无法满足学者多元化、个性化的知识需求和专业性、及时性的服务需求，以丁香园、小木虫等为代表的社交类学术移动应用程序（以下简称“社交类学术 App”）已成为科研用户获取知识资源、进行学术交流的新途径[1]。作为依托用户交互行为存在的新兴平台，用户的持续使用是其运营发展的关键，因此，如何识别用户动态行为演化并根据不同类型用户特征提供精准服务，成为平台运营面临的重大挑战。

用户画像作为勾画目标用户、提高决策效率的有效工具，现已在多领域得到广泛应用。R. J. Holden 等从年龄、性别、经济背景等维度构建了老年用户健康角色模型[2]；M. Trusov 等通过分析个人资料和行为数据描绘消费者兴趣偏好[3]，这种根据用户基本属性与行为特征的画像对识别典型群体有一定指导意义，但画像粒度粗糙，“千人一面”的角色模型无法挖掘出用户深层需求。因此部分学者尝试从用户个体小数据层面构建更为全面精准的画像模型：陈臣等通过采集图书馆用户小数据构建了面向读者个性化服务的精准画像模型[4]；孙丹霞等认为依托小数据可以根据用户全方位行为特征及情境感知构建生动全面的“用户自画像”[5]。引入全面表征用户个体特征的小数据可有效细化画像粒度，但目前基于小数据构建的画像多为采集某一时间节点数据的即时性画像，生成的角色模型是一个相对静止的状态，只能展现用户当时当下的行为特征，无法对其未来行为趋势作出合理推断。此外，面对数据激增，即时性静态画像多采用重复迭代的方式重新刻画用户全貌，这种方式效率低下，且未充分利用之前的画像信息，实用价值有待提升。为数不多的用户动态画像的探索性思想多面向图书馆领域[6-7]，尚缺乏在社交类学术 App 情境下的应用尝试。

综上所述，构建细粒度的用户动态画像是完整展现用户概念全貌、实时洞察用户行为演化、提高平台精准服务水平的重要方式。因此，本文以社交类学术 App 为研究对象，基于与社交类学术 App 用户强相关、高价值、全方位的小数据构建突破用户表层行为差异的动态画像模型。在描述用户行为特征的基础上探求影响其行为的动因与触发点，勾勒出具有稳定性、持续性、动态性的用户角色，以便平台运营者深刻理解用户行为需求及演化趋势，从而预见性地提出个性化运营

* 本文系国家社会科学基金项目“大数据驱动下学术新媒体知识聚合及创新服务研究”（项目编号 18BTQ085）研究成果之一。
作者简介：张莉曼（ORCID：0000－0002－0770－3708），博士研究生，E-mail：326671265@qq.com；张向先（ORCID：0000－0003－3186－2677），教授，博士，博士生导师；吴雅威（ORCID：0000－0001－9703－8731），博士研究生；郭顺利（ORCID：0000－0002－3155－9937），讲师，博士。
收稿日期：2019－06－25 修回日期：2019－09－19 本文起止页码：50－59 本文责任编辑：易飞

策略，助力产品与用户的精准对接。

## 1　相关研究与问题的提出

### 1.1　小数据

小数据概念最初由康奈尔大学的 D. Eestrin 教授发现并提出，认为可通过对用户日常行为全方位数据的追踪，动态监控用户的健康变化[8]。目前学界对小数据暂无明确统一的定义，但均认可小数据是以人或团队为中心的全方位、多层次行为模式和情境感知的全部数据集合[9]。随着时间的推移，这些数据集合不断丰富，为动态挖掘用户需求偏好及行为规律提供了有力支撑。当前关于小数据的研究集中于个性化推荐[10]及精准服务[6]、兴趣发现与预测[11]以及小数据融合的理论探讨[12]等方面，尚缺乏依托小数据构建学术 App 画像模型的概念构想。

通过对小数据相关研究及应用的梳理，本文认为小数据具有以下典型特征：①用户中心性。与关注宏观总体的大数据有别，小数据是围绕用户展开的、能够展示用户真实内在的个体化数据，价值密度更高，为精准描述用户全方位行为特征及概念全貌提供支撑；②多维复杂性。与大数据相比，小数据更注重对个体全景数据不间断、多维度、深层次采集及情景因素的关联，数据来源更为广泛，数据类型更为多样，因此在进行小数据处理时需融合多元数据处理方法；③关注因果关系。大数据着眼于数据之间表层相关性描述，不探究影响数据相关性的深层原因，而小数据不仅关注数据相关性表现，也注重揭示数据相关关系的驱动因素。此外，小数据同样具有大数据的价值性、动态性、快速性等特征，可视为大数据的补充和延伸，因此可充分借鉴大数据相关技术进行小数据处理与利用。

### 1.2　社交类学术 App 用户动态画像

作为兼顾学术性与社交性的一体化平台，社交类学术 App 是指安装在移动智能终端上为用户提供学术资源或交流平台的应用程序[13]。当前，由于移动智能终端的普及和 App 的广泛使用，对 App 用户进行画像描摹引起了多领域学者的关注。例如纪庆楠通过建立用户画像与情绪波动图获取智能公交 App 用户需求痛点[14]；李大伟等根据用户画像与协同过滤算法设计了图书推荐 App 个性化推荐模式[15]；韩张俊杰以资讯类 App 为切入点构建用户画像，利用聚类算法与关联规则划分用户群体并挖掘群体特征，旨在优化精准服务模式[16]。上述研究为学者及运营人员以用户画像的方式定位群体需求、实现个性化推荐、践行精准营销提供了参考借鉴，但社交类学术 App 尚属新生事物，相关研究成果十分匮乏，以往研究主要通过理论性探讨或问卷调查法对其技术开发方法[17]、用户使用[18]或采纳的影响因素[13]进行探讨，缺乏数据驱动下以用户画像的方式细分用户群体、把握用户需求演化的探索尝试。

传统的用户画像刻画时仅采集某一时间节点上的数据，即根据用户的行为特征、生活习惯等数据标签，抽象出一个能静态展示用户现实及历史全貌的模型[16]。社交类学术 App 用户动态画像指在刻画用户概念全貌的基础上引入时间片段，选用科学的方法，动态、持续地勾勒用户与平台交互过程中行为轨迹的发展趋势。从这个意义上看，静态画像可看作动态画像在某一时间节点上的定格描述。目前已有部分学者对动态画像的构想进行积极尝试：刘勇等根据用户历史交互数据预测其兴趣变化趋势并进行动态化推荐[19]，但忽视了用户的主观易变性，基于历史数据的推荐不一定能满足用户现在或未来的需求；王益成等认为可通过采集用户行为大数据构建行为标签库，然后根据用户反馈不断修正画像模型[20]，仍只关注了表层行为，未考虑行为背后的深层动机，画像粒度较粗。张慧敏探讨了生活方式转型背景下动态用户画像的必要性及构成维度，但侧重于分析构建动态画像对交互设计师的要求[21]，同时未考虑画像随时间变化的演化规律。由此可知，现有的用户动态画像方法存在画像粒度粗糙、时效性差的弊端，针对社交类学术 App 特点的用户动态画像研究十分匮乏。

### 1.3　基于小数据的社交类学术 App 用户动态画像的提出

社交类学术 App 用户多为受教育程度较高、有专业研究领域的科研人员，与一般 App 相比需求特征更为明显，且依托于移动终端的便捷性，有助于满足用户即时性、情景性的学术及社交需求。因此设计画像标签体系时应全方位、多维度、深层次挖掘用户行为特征及深层驱动因素，同时考虑用户在特定情境下的状态及需求。此外，动态画像要求数据具有延续性，能够持续稳定地揭露用户的特征趋势。根据小数据概念可知，社交类学术 App 用户小数据是指围绕用户使用平台全过程展开的各类数据集合，能够真实、全面表征用户行为、动机、使用情景等细粒度特征，并且通过对个体用户行为特征长时间的监测获得，可以满足构建动态画像的数据延续性要求。

与基于大数据构建用户动态画像相比，基于小数据的社交类学术 App 用户动态画像具有以下典型特

征:①精准性。大数据主要来源于大量用户的行为活动,描绘的画像聚焦于大量用户同一类型的特征,小数据来源于个体用户的各类活动,关注的是单一用户的多维特征,基于小数据的画像模型更能精准展示用户概念全貌;②深层次。常见的大数据画像多通过采集用户基本属性及行为数据构建而成,适用于从行为层面对用户群体进行初步划分的研究场景,而小数据画像还强调突出行为驱动因素,更利于用户动态化行为规律的深度挖掘与个性化运营策略的及时制定;③实用性。大数据的数据体量庞大,价值密度低,其中包含大量会干扰画像精度的无效数据,导致基于大数据的动态画像效率低下。而小数据围绕针对性强的个体用户展开,数据体量适中,并且小数据采集过程在相对封闭的环境中进行,更易与用户建立良好的沟通机制,减少用户隐私泄露的顾虑,获取到高价值数据。因此,基于小数据的动态画像模型实用价值更高。

综上所述,基于小数据的动态画像完全契合社交类学术 App 平台高效追踪用户特征趋势、及时制定精准营销策略的战略目标。因此,本文从小数据视角入手,探究社交类学术 App 用户动态画像模型构建的思路方法。首先明确基于小数据的社交类学术 App 用户动态画像构建总体流程,然后以勒温场动力理论为依据确定动态画像的构建维度,并据此提出相关小数据的采集及处理方法。接下来探究勾勒动态画像的适用方法,最后形成基于小数据的社交类学术 App 用户动态画像框架模型,为数据驱动情境下社交类学术 App 平台精准运营提供新的研究视角和思路参考。

## 2　基于小数据的社交类学术 App 用户动态画像总体设计

典型的用户画像构建方法有 A. Cooper 的"七步人物角色法"和 L. Nielsen 的"十步人物角色法",这两个方法在流程上可概括为获取标签数据、细分用户群体、建立并丰富用户画像三个环节[22]。根据社交类学术 App 用户动态画像的概念可知,构建时需在传统画像基础上引入时间片段,根据相邻时间段簇族的迁移关系发现用户行为的动态轨迹。此外,本文提出的基于小数据的动态画像模型更加强调标签体系的针对性和立体化。因此,基于小数据的社交类学术 App 用户动态画像模型流程应分为 4 个环节,如图 1 所示:

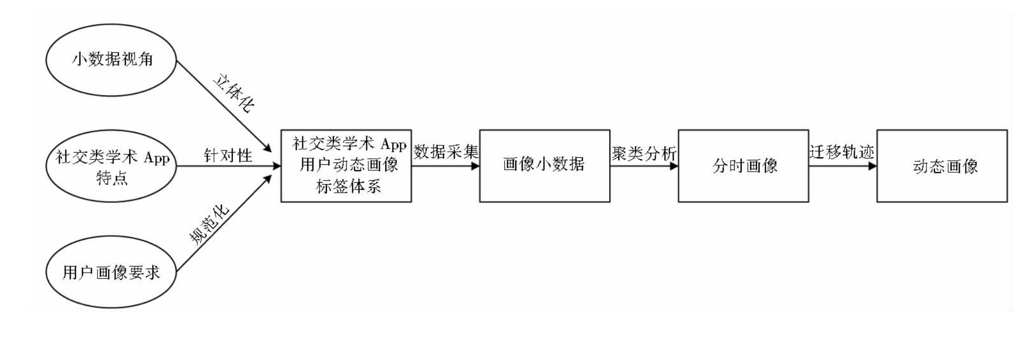



图 1　基于小数据的社交类学术 App 用户动态画像构建流程

Step1:从小数据视角出发,结合社交类学术 App 特点设计包含用户行为深层驱动因素的立体化维度标签;

Step2:根据维度标签采集用户小数据并进行预处理;

Step3:根据用户小数据中的时间信息划分数据,利用聚类算法对特定时间片段内的数据聚类,将用户划分为不同群体,构建出分时画像,并将其储存在数据库中;

Step4:根据分时画像,确定各个时段内的类簇中心(类簇中心可视为该群体用户的典型代表)。挖掘并分析相邻时段内类簇中心的动态迁移关系,追踪用户行为变化轨迹,实现用户动态画像的描摹。

上述设计流程既参考了传统用户画像的基本环节,又融合了本文提出的基于小数据的动态画像的特殊性及要求,实用价值与创新作用得以保证。

## 3　社交类学术 App 用户动态画像标签体系与小数据获取

### 3.1　标签维度确定的理论依据

用户的心理活动是支撑行为产生的内在因素,任何行为发生前都会受到一定的意图驱使[23],社会心理学家勒温提出了场动力理论,用于分析支撑个体行为产生的驱动力和行为变化过程。场动力理论包括场论和动力论,其中,场论将"场"定义为个体与环境相互

依存的整体形态，也称为个体生活空间(LS)。个体的心理及行为总是在这个空间内发生并移动，用函数公式可表示为：

$$B = f(P * E) = f(LS) \quad 公式(1)$$

B 代表外化的行为表现，P 代表个体内在需求，E 代表心理环境，即对个体内在需求产生刺激作用的情景，f 为个体与环境相互作用的函数[24]。因此，场论认为个体行为是主体与情境交互作用的结果。动力论提出个体心理或行为的动力源于个体与情景交互过程中产生的紧张感。即当个体需求未得到满足时，其心理便会处于紧张状态，驱动行为产生以缓解或消除心理张力。此外，个体的心理目标也是驱动行为产生的重要因素。根据场动力理论可知，社交类学术 App 用户行为会受到内在需求和外在情境的双重驱动。具体来说，用户对社交类学术 App 往往有一个直观、基础的内在需求，例如满足查找文献的需要、进行科研合作的需要等。用户所处情景或与其他用户的交互情景也会催生需求的产生，当用户发现这些需求无法得到满足时便会产生心理张力，从而进行一系列行为活动企图消除心理紧张感，如在平台查阅、下载所需知识或提问、收藏等表达个人诉求与兴趣偏好。因此，社交类学术 App 用户行为、内在需求与外在情境之间呈现出动态交互关系，为本文深入分析用户行为驱动过程并确定动态画像维度提供了理论依据。

## 3.2 标签体系的构成及小数据采集

根据场动力理论与社交类学术 App 特征，本文认为驱动社交类学术 App 用户行为的因素包括用户的价值取向、认知能力、情景特征和社交关系。结合用户自然属性与行为偏好两个画像基本因素，构建出包括 6 个维度的画像体系，如图 2 所示。各维度内在关系为：用户的自然属性与行为偏好是勾勒画像的基础框架，价值取向、认知能力、情景特征和社交关系驱动了行为的产生，其中价值取向和认知能力属于用户自身驱动，即内驱力 P，情景和社交为外界情境的刺激因素，即诱因 E。行为偏好是自然属性、价值取向、认知能力、情景特征、社交关系综合作用的外化体现。

(1)自然属性。具有持久稳定特征的自然属性是催生用户行为变化的基础[25]。其中，性别与年龄是群体行为、偏好及需求趋向的影响因素；社交类学术 App 以知识的提供和交流为目的，用户可根据自身兴趣需求获取相应内容及服务，因此也应考虑用户的教育程度和专业领域。

(2)行为偏好。用户行为偏好指用户对某一事物

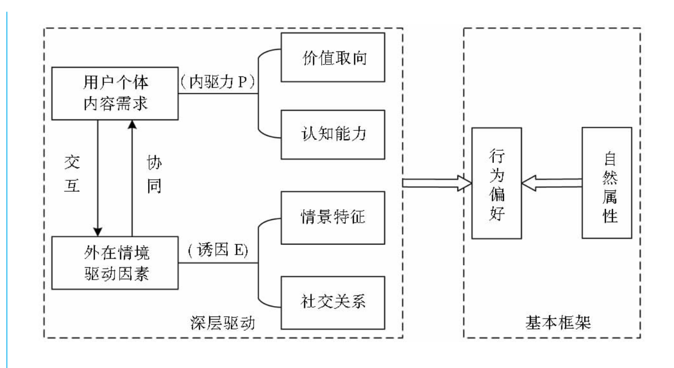



图 2 社交类学术 App 用户动态画像维度体系

的倾向及关注程度，是需求的外化体现[22]。把握用户行为偏好需从直观需求和兴趣倾向两方面考虑。用户往往通过搜索、浏览、点击、下载获取所需知识，或由发帖、意见反馈直观表达自身的知识诉求，而关注感兴趣的用户、收藏价值内容则是兴趣倾向的体现。

(3)价值取向。自我差异论认为，理想自我是期望中个体所应具备的能力特征，代表了对自身未来状态的愿景，现实自我指当下个体具备的能力特征，二者之间的差距驱使用户产生能够不断弥合差距的行为[26]。科研用户使用社交类学术 App 旨在通过解决科研问题或科研互助等提高自身能力，不断接近期待的自我。因此，价值取向是用户行为产生的动力源泉，可从用户愿景与自我评估两方面考察。

(4)认知能力。认知能力指用户对信息内容进行识别、加工并有效运用的能力，认知能力的不同驱使用户产生差异化行为[27]。认知能力一方面与用户素质和文化水平相关，另一方面与平台交互程度相关。用户认知能力在与平台的交互过程中逐步提高，对平台的价值贡献越来越大。因此，可从用户自身情况和平台贡献价值两方面考虑其认知能力。其中，等级、认证身份是用户自身水平的体现，被关注数、被收藏数、被点赞数则是对平台贡献价值的反映。

(5)情景特征。不同情景下用户需求的变化会导致行为偏好发生改变倾向[28]。情景是动态的、连续的，情景片段的链接形成了科研用户的生活轨迹，因此，分析情景因素是构建发展性用户动态画像、及时响应用户需求的必然要求。根据情景分类[29]及研究对象特征，本文认为可从时间情景、位置情景、用户情景以及设备情景四方面感知情景特征。

(6)社交关系。社交类学术 App 在满足用户知识需求的同时还兼具社交功能，鼓励用户积极参与知识交流、共享与创新[25]。平台上的用户在与他人关注、讨论、分享过程中可进一步挖掘自身潜在需求与兴趣，

进而影响使用 App 的行为模式。因此社交关系可视为驱动用户行为模式的群体网络情境，用户加入的群组、关注的人数与互动数据是动态性深入追踪用户行为轨迹的重要标签。

用户的基本信息、社交关系数据存储于 App 管理后台；行为偏好可从存储于用户日志上的行为数据和用户生成内容中获取；价值取向可挖掘用户生成内容获得，也可通过问卷、访谈等调研方式得到，例如平台发放定向电子问卷来了解用户愿景及当下需求；认知能力可由用户提交的认证资料及交互数据分析获得；对于情景特征的感知主要依靠传感器、定位系统和智能穿戴设备；社交关系利用日志挖掘及社交网络分析法得出。基于小数据的社交类学术 App 用户动态画像标签体系及小数据采集方法如表 1 所示：

表 1　社交类学术 App 用户动态画像标签及小数据采集

| 维度 | 标签 | 标签解释说明 | 小数据来源与采集方法 |
|---|---|---|---|
| 自然属性 | 用户性别 | 人口统计学特征，分为“男”、“女” | 用户注册时所提交的个人信息；可在 App 数据管理后台上直接获取 |
| | 用户年龄 | 人口统计学特征，可按年龄段划分 | |
| | 教育程度 | 用户注册时所填的教育及学历水平… | |
| | 专业领域 | 用户注册时所填的研究或工作领域 | |
| 行为偏好 | 搜索行为 | 用户在 App 上搜索内容或学者的行为 | 用户在平台上留下的使用痕迹和生成的文本内容，保存在平台上的用户日志中；利用网络爬虫、日志挖掘技术或埋点技术动态追踪用户行为，通过数据挖掘方法识别用户兴趣偏好 |
| | 浏览行为 | 用户在 App 上浏览相关资源或社群交流内容等 | |
| | 点击行为 | 用户点击链接或图片视频等内容以进行详细查看 | |
| | 下载行为 | 用户将需求资源下载下来以便保存的行为 | |
| | 反馈行为 | 用户将需求、意见建议等反馈给其他用户或平台 | |
| | 关注话题 | 用户自主在平台上关注的感兴趣的话题 | |
| | 收藏内容 | 用户自主在平台上收藏的感兴趣或有价值的内容 | |
| | 发帖内容 | 用户在平台上的提问发帖或回复他人的帖子 | |
| 价值取向 | 用户愿景 | 用户期待使用 App 能够达到的理想自我的水平 | 用户生成内容、问卷或访谈；文本挖掘或调研 |
| | 自我评估 | 用户对现实自我水平及状态的评估 | |
| 认知能力 | 用户等级 | 用户在平台上的等级，与平台交互越深入等级越高 | 用户注册时认证的个人信息及交互过程中产生的数据：可在 App 数据管理后台上直接获取 |
| | 认证身份 | 由用户个人提交并经平台认证的身份信息 | |
| | 被关注数 | 用户被他人关注的数量，是用户对平台价值的体现 | |
| | 被收藏数 | 发表内容被他人收藏的数量，是用户对平台价值的体现 | |
| | 被点赞数 | 用户发表内容被他人点赞的数量，是用户对平台价值的体现 | |
| 情景特征 | 用户情景 | 用户心理状态，可分为任务情景、休闲情景、其它情景 | 平台监管系统自动感知并记录；由传感器、定位系统、智能穿戴设备等获取用户使用的时空物理特征和心理状态 |
| | 时间情景 | 用户使用 App 时段，可分为晨起、上午、中午、下午、睡前 | |
| | 位置情景 | 用户使用 App 时所处的空间地理位置 | |
| | 设备情景 | 主要包括硬件信息（屏幕大小）和网络信息（网络状态） | |
| 社交关系 | 群组数量 | 用户加入或参与的相关讨论群组数量 | 用户在平台上的行为痕迹；通过日志挖掘、社交网络分析获得用户社交网络结构 |
| | 关注人数 | 用户自主关注的平台上其他用户的数量 | |
| | 互动数据 | 用户间的讨论、分享、合作等交互数据 | |



### 3.3　小数据处理

以往用户画像应用的数据多为基本属性数据与行为数据，可通过编码方式进行数值转化或简单处理后直接用于实验分析[3]，但本文构建的画像在上述基础上还需融合评论或发帖文本（即内容特征）。文献[30]提出了一种基于行为 - 内容融合模型的画像方法，首先将用户发表文本进行拼接，然后进行深度用户表示学习，再通过聚类获得类别标签，将其作为一个特征加入行为特征中共同作用于画像的勾勒。这种方法为多源数据处理提供了一定的参考，但未考虑文本内容的主题特征。社交类学术 App 的专业性、领域性突出，用户发表、回复帖子或关注的主题往往与自身特定领域相关，这种特点使从文本内容中挖掘分析出用户的兴趣领域成为可能。因此，本文提出了一种基于 LDA 主题模型的文本数据建模方法，见图 3。

Step1：采集文本内容，按用户 ID 逐条拼接并存储为文本文档，清洗后导入领域词表进行中文分词、停用词过滤，将原始文本切割成以特征词为单元的序列，再利用 TF-IDF 从词频和重要性两个角度计算特征词权重，保留重要特征词；

Step2：利用 LDA（Latent Dirichlet Allocation，隐含狄利克雷函数）主题模型中简单易用的 Gibbs 采样挖

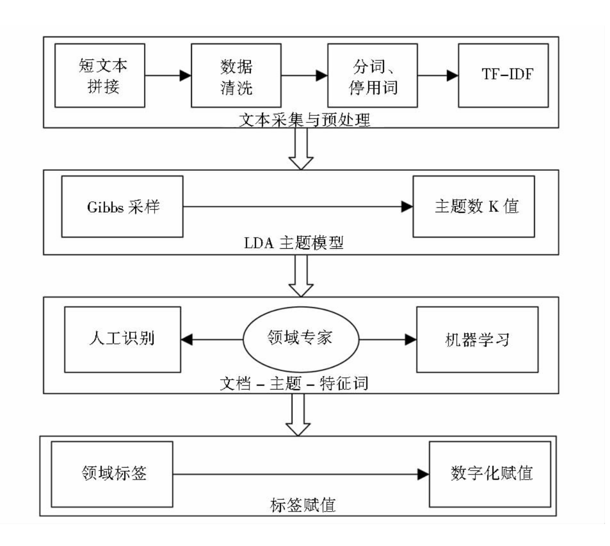



图 3 基于 LDA 主题模型的非结构化文本建模步骤

掘文本隐含的主题,即根据统计学思想将复杂的文本投影到潜在的主题空间,得到"文档 - 主题"分布 $\theta$ 和"主题 - 特征词"的分布 $\varphi$ [31]如图 4 所示。主题个数可根据困惑度评价法确定[32]:

$$perplexity(D) = \exp\left[ -\frac{\sum_{d=1}^{M} \log(P(W)_d)}{\sum_{d=1}^{M} N_d} \right]$$ 公式(2)

$$p(w) = p(z \mid d) * p(w \mid z)$$ 公式(3)

其中,M 为文本数,D 为测试集文档,Nd 为文档 d 中出现的所有词总数,p(w)为测试集中每个词出现的概率。p(z|d)为文档中各个主题出现的概率,p(w|z)表示某主题下每个特征词出现的概率[32]。困惑度随主题数 K 值的增加而下降,下降趋势趋于平缓时的 K 值为最佳个数。然后根据相关性对主题下的词排序,取前 N 个作为特征词,形成 N 个"主题 - 特征词"矩阵。

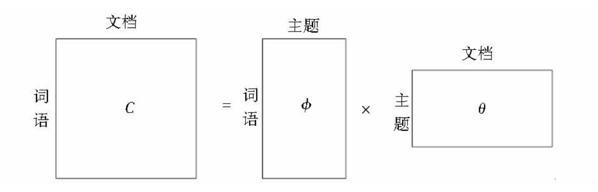



图 4 "文档 - 主题 - 特征词"矩阵

Step3:得到"文档 - 主题 - 特征词"矩阵后,若样本较少,可人工识别各主题词簇,识别时应综合考虑主题下词语的分布情况和语义关系。由于社交类学术 App 多为行业垂直类平台,例如面向医疗领域的丁香园 App 和面向经管领域的经管之家 App,因此在识别主题特征词时可咨询相关领域专家,为各文档设置一个能够概括其特征的领域标签。若样本较多,可以采用包括但不限于 KNN(k-Nearest Neighbor)机器学习中的分类算法进行文档标签的匹配。

Step4:获得各文档领域标签后,对其进行数字化赋值,例如将"内科""外科""中医"三个标签分别赋值{1,2,3},最终实现从用户生成内容中识别用户兴趣领域并将其转化为数值的目的。

上述方法既可对文本数据进行量化,又能从主题层面对文本特征进行概括揭示,兼顾了实用性与科学性,实现了将文本数据与其他数据共同作用于画像描绘的研究目标,能够最大限度避免画像失真。

## 4 关键技术介绍与框架模型的形成

### 4.1 动态画像构建的关键技术

运用聚类算法将大规模用户划分为几个典型群体,可在数据驱动环境下高效把握用户核心特征及需求[22]。这一思想已得到广泛应用:如陈添源采用 K-means 聚类描绘移动图书馆差异化用户群体[33];陈娟等采用层次聚类识别出知乎平台三类典型群体[34]。社交类学术 App 用户量较大,适合通过聚类方法挖掘群体画像特征,但不同于以往面向静态数据的聚类,本文提出的动态画像引入了时间片段,通过识别相邻时间段簇族的迁移关系来实现用户动态轨迹的描摹。目前鲜有研究对基于时间序列数据的动态画像方法进行探讨,但学者广泛认同对时间序列数据聚类需考虑两个核心问题:①特定时段内的聚类结果应充分反映出该时段内数据的特征;②不同时段的聚类结果在时间轴上呈现出一定的连续性,即相邻时段的簇族是平滑演化的[35]。D. Chakrabarti 等于 2006 年首次提出演化聚类思想,并将其应用到 K-means 算法上得到演化 K-means 聚类,用以解决时间序列数据聚类准确性与连续性问题[36],在此基础上,王富鹏考虑了历史数据对当前时刻聚类结果的影响,并将其应用于金融股票行情轨迹的趋势分析,帮助股民实时了解股市变化[37]。由于本文的研究目标、数据结构与上述研究相似,因此本文认为可充分借鉴演化聚类思想动态挖掘社交类学术 App 用户演化行为,步骤如下:

Step1:划分合理的时间窗口。将采集的数据按照一定的时间段 t 分割成多组数据,或按照时间段多次采集数据,获得不同时间段下的多组数据。对于 t 值可根据 App 产品开发或迭代的周期确定,也可参考损失函数确定[35],即:$Cost = \alpha \cdot CS + (1 - \alpha) \cdot CT$。其中,CS 为快照损失(cost of snapshot),值越大表明当前时段内聚类效果越差;CT 为时间损失(cost of temporal-

ity),时间损失越大证明相邻时段的平滑性越差。α 为用户对上述两个值的权衡系数,α 越大说明用户越重视片段内聚类质量,α 越小则相反。

Step2:构造用户分时画像。这一过程借鉴文献[37]提出的演化 K-means 算法实现,即在传统 k-means 算法聚类中心点选择的基础上融入对历史时间片段内聚类中心权重的考虑,公式为:

$$C_j^t \leftarrow \alpha \times E_{x \in closest(j)}) \mid (x) + (1-\alpha) \times \sum_{i=1}^{t} (f(t,t-i) \cdot C_j^{t-i}) \quad \text{公式(4)}$$

其中,$C_j^t$ 代表时段 t 内的第 j 个聚类中心点,E 为期望运算,$closest(j)$表示距离中心点 j 最近的样本点,$f(t,t-i)$指 $t-i$ 时段聚类中心的权重。α 仍为快照损失与时间损失的权衡系数。此外,选择初始中心点时并无历史数据可以参考,且 K-means 算法存在需人为指定簇族个数与随机选择初始中心点的弊端,因此可将 Canopy 算法作为 K-means 聚类之前的先验簇族依据[38],同时确定 Silhouette 系数高且各组样本分布合理时的点为初始中心点[39],再完成对时间序列数据的分组聚类。

Step3:识别演化轨迹。这一环节需分析相邻时段内簇族的映射关系,即判断簇族的相似度,可通过计算条件概率得到簇族权重的方法实现,具体方法为:已知两个相邻时刻 $t_i$ 和 $t_{i+1}$ 的聚类结果分别是 Qi 和 $Q_{i+1}$,$Q_i$ 中的簇族 $C_m(t_i)$与 $Q_{i+1}$中的簇族 $C_u(t_{i+1})$的权重计算公式[37]为:

$$\text{Weight}(C_m(t_i), C_u(t_{i+1})) = P(X \in C_u(t_{i+1}) \mid X \in C_m(t_i)) = \frac{\sum P(x \in C_m(t_i) \cap C_u(t_{i+1}))}{\sum P(x \in C_m(t_i))} \quad \text{公式(5)}$$

其中,$P(X \in C_u(t_{i+1}) \mid X \in C_m(t_i))$表示样本点在属于 $C_m(t_i)$条件下属于 $C_u(t_{i+1})$的概率,$P(x \in C_m(t_i) \cap C_u(t_{i+1}))$样本点属于 $C_m(t_i) \cap C_u(t_{i+1})$的概率,$P(x \in C_m(t_i)$表示样本点属于 $C_m(t_i)$的概率。由于时段内数据的不平衡性,簇族一般会有 7 种演化状态,如图 5 所示,分别为:出现一个新的群体;一个群体分裂为两个或两个以上群体;两个或两个以上群体合并为一个群体;一个群体在下一阶段消失;某群体内的用户数量在下一阶段增加;某群体内的用户数量在下一阶段减少;某群体在相邻阶段未发生任何变化。在识别演化行为之前需要事先设定簇族演化的临界值,M OLIVEIRA 等人提出的 MEC 框架(Monitor of the Evolution of Clusters over time)中通过定义 survival 的参数 τ 和 split 的参数 φ 来识别各个演化行为[40],王富鹏在此基础上引入参数 μ 作为 grow 和 decline 行为的临界值以识别上述 7 种演化行为。三个参数的阈值均为[0,1],具体取值应根据实际应用场景,通过反复迭代实验确定[37]。

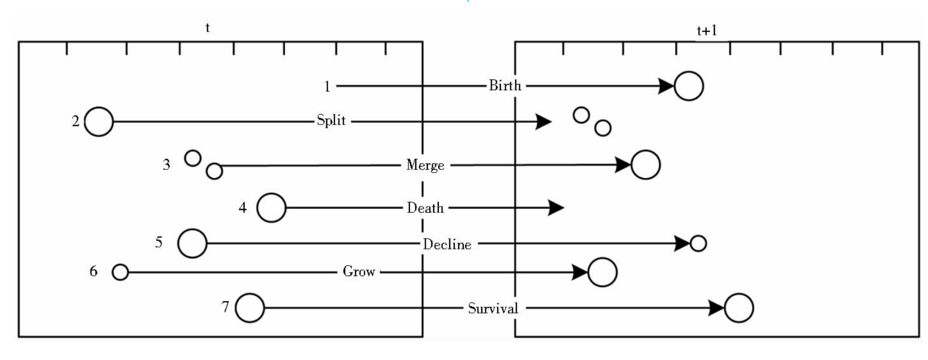



图 5 簇族演化示意

社交类学术 App 用户的行为习惯往往具有延续性,演化聚类考虑了历史数据对当前聚类的影响作用,利用演化聚类构建动态画像更贴合社交类学术 App 用户实际状态,且融入了 Canopy 算法的演化 K-means 算法鲁棒性更强[38]。此外,随着用户与平台交互程度的加深,用户所属群体是一个动态变化的过程,这与簇族演化过程分析是一致的。因此,利用演化 K-means 算法实现社交类学术 App 动态画像理论上具有极强的可行性,且相关应用研究[37]为其提供了实践支撑。

### 4.2 基于小数据的社交类学术 App 用户动态画像框架模型

由于构建目标与应用场景不同,画像模型的层次结构与具体构建方法也有所差异。Y. Kritikou 等认为模型应包括监控层、建模层、适应层三个层次[41];许鹏程等从数据采集、处理、存储、挖掘、呈现及应用 6 个层面构建数字图书馆用户画像框架模型[25]。借鉴上述研究,本文认为基于小数据的社交类学术 App 动态画像实质是在标签体系的基础上充分采集与用户强相关的小数据,通过引入时间窗口构建用户分时画像并通

过簇族迁移识别用户动态轨迹。因此，框架模型应包括小数据采集处理层、分时画像构建层、动态画像形成层三个层次，如图 6 所示：

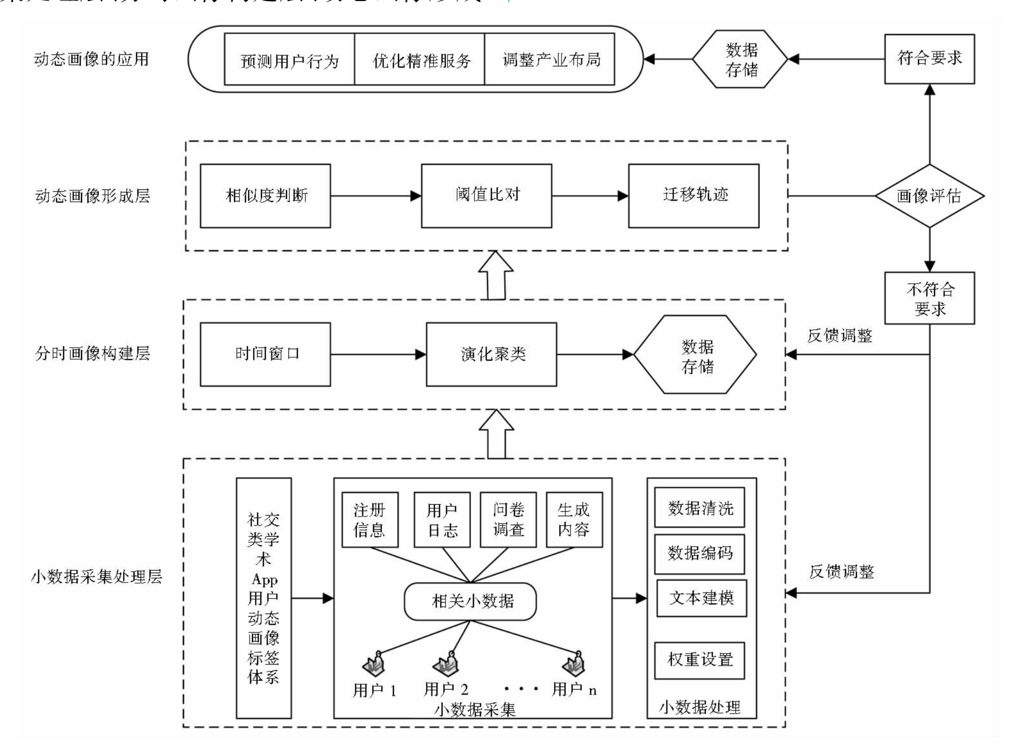



图 6　基于小数据的社交类学术 App 用户动态画像框架模型

#### 4.2.1　小数据采集处理层

小数据采集处理层根据构建的标签体系，综合运用网络爬虫、日志挖掘等技术与能够获取用户深层特征的调研等方式获得相关小数据并进行预处理，转化为满足画像需要的数据形式。由于勾勒画像的各维度标签与画像结果的相关度不一，其相关程度与动态画像的具体应用情境有关。因此应根据实际需要调整标签权重，确保画像结果的科学决策价值。

#### 4.2.2　分时画像构建层

构建用户分时画像是识别其行为轨迹的前提。首先按照数据的时间信息将其划分到特定时段中，形成时间序列数据，然后对不同时段内的数据分别聚类，即将特定时段内具有相似特征的用户划分到同一类簇，并确定各类簇中心点，最后将分时画像结果自动存储于数据库中，为下一步识别用户动态行为轨迹做铺垫。

#### 4.2.3　动态画像形成层

形成动态画像需通过相似度判断、阈值比对、动态轨迹识别三部分完成。首先计算相邻时段簇族的权重，然后将所得权重与事先设置的阈值对比，判断簇族间的新生、合并、分裂、消亡等迁移关系，识别出用户动态迁移轨迹。为了保证画像的准确性和实用性，对画像结果进行评估，若能满足平台决策要求，则将画像结果存储以便利用；若不能满足，则通过反馈调整各层级的组织内容及流程结构，构建出真实可靠、适用性强的动态画像。

上述三个层次形成了一个闭环过程，既符合画像模型的一般流程，又能实现基于小数据描摹社交类学术 App 用户动态画像的研究目标，并且画像评估模块考虑了画像的实际应用价值，科学性更强。

### 4.3　基于小数据的社交类学术 App 用户动态画像的作用

#### 4.3.1　预测用户行为

对平台现有用户进行动态画像，可通过观察其动态趋势预测下一阶段的行为特征，有效改善了传统画像方法的滞后性弊端；同时，通过特征对比或中心点距离计算将新用户定位到分时画像数据库中相似度最高的簇族，然后根据轨迹模型预测出用户在下一时段大概率会发生的行为，一定程度上可以解决新用户的冷启动问题。

#### 4.3.2　优化精准服务

本文构建的社交类学术 App 用户动态画像模型以小数据作为画像数据支撑，既考虑了用户的表层行为特征，又融合了深层行为驱动因素，描摹出的画像更贴合用户的实际状态与全面诉求。社交类学术 App 的运

营者可据此细化用户群体,针对群体特征推荐适配性内容,实现用户与资源的精准匹配,从而优化平台精准服务水平,提高决策效率及能力。

#### 4.3.3　调整产品布局

用户行为轨迹的识别与追踪使 App 平台实时、动态把握用户需求成为了可能。构建动态画像有助于为用户量身定制覆盖用户与平台交互全过程的产品及服务,同时也可根据用户需求的演化辅助平台具有前瞻性地调整产品布局,优化平台结构,动态性对接用户个性化需求,不断提高平台的市场竞争力。

## 5　结语

作为大数据的延续与补充,小数据在全方位、深层次表征个体用户行为模式及情境因素方面具有极强的优越性。同时,无线传感器技术、智能可穿戴设备以及监控、定位技术的成熟与普及为小数据实时获得提供了技术支撑。鉴于此,本文提出基于小数据描摹社交类学术 App 用户画像的思路,并探究了运用演化聚类及簇族迁移实现动态画像的可行性,构建出基于小数据的社交类学术 App 用户动态画像框架模型并阐述其在预测行为、优化服务、调整布局方面的作用。为社交类学术 App 平台运营者精准洞察用户动态化需求、及时制定适配的资源服务策略提供新的视角和参考思路。

由于目前基于小数据视角构建社交类学术 App 用户动态画像的研究十分匮乏,本文旨在提出一个全面、系统的框架并进行可行性分析,丰富并拓展小数据动态画像的理论研究体系,为小数据的应用与用户画像的创新突破提供新的视角。限于篇幅,无法在一篇文章中全面展示相关实证过程,后续将在本文基础上对小数据采集处理及动态画像的构建进行实证分析,深入探讨模型的应用价值与泛化作用。



**参考文献:**

[ 1 ] 耿斌,孙建军. 在线学术社交平台的用户行为研究——以 ResearchGate 平台南京大学用户为例[J]. 图书与情报,2017,61(5):47 –53.
[ 2 ] HOLDEN R J, KULANTHAIVEL A, PURKAYASTHA S, et al. Know thy eHealth user: development of biopsychosocial personas from a study of older adults with heart failure[J]. International journal of medical informatics, 2017, 108(12):158 –167.
[ 3 ] TRUSOV M, MA L, JAMAL Z. Crumbs of the cookie: user profiling in customer-base analysis and behavioral targeting [J]. Marketing science, 2016,35(3):405 –426.
[ 4 ] 陈臣,马晓亭. 基于小数据的图书馆用户精准画像研究[J]. 情报资料工作, 2018,40(5):57 –61.
[ 5 ] 孙丹霞,王伟军,姜毅. 基于用户小数据的嵌入式学科服务研究[J]. 图书馆工作与研究,2019(4):84 –90.
[ 6 ] 刁羽,畅佩. 面向小数据的图书馆精准创客服务研究[J]. 图书馆理论与实践,2018(5):109 –112.
[ 7 ] 王欣,张冬梅. 大数据环境下基于高校读者小数据的图书馆个性化智能服务研究[J]. 情报理论与实践,2018,41(2):132 –137.
[ 8 ] ESTRIN D. Small data, where n = me [J]. Communications of the ACM, 2014, 57(4):32 –34.
[ 9 ] 陈臣,马晓亭. 基于小数据的图书馆用户精准画像研究[J]. 情报资料工作,2018(5):57 –61.
[10] HSIEH C K, YANG L, WEI H, et al. Immersive recommendation: news and event recommendations using personal digital traces [C]//Proceedings of the 25th international conference on World Wide Web. Montreal: ACM, 2016:51 –62.
[11] 陈臣,李强. 基于小数据决策的读者兴趣发现与预测[J]. 情报科学,2017,35(5):75 –80.
[12] 李立睿,邓仲华. “互联网 +”背景下科研用户的小数据融合研究[J]. 图书情报工作,2016,60(6):58 –63.
[13] 张晓丹,江洪,王可慧. 学术 App 用户采纳意愿影响因素实证研究[J]. 图书情报工作,2018,62(18):90 –101.
[14] 纪庆楠. 基于用户体验的智能公交 APP 交互设计研究[D]. 西安:西安理工大学,2017.
[15] 李大伟,杜洪波,周孝林,等. 基于“用户画像”挖掘的图书推荐 App 设计[J]. 软件,2018,39(5):35 –37.
[16] 韩张俊杰. 基于数据分析的资讯类 App 用户画像设计与应用[D]. 北京:中国科学院大学,2017.
[17] 郭伟. 基于云平台的科技期刊 APP 开发方法研究——以“长白山学术汇”为例[J]. 中国科技期刊研究,2018,29(5):485 –490.
[18] 王慧. 学术期刊 APP 使用的影响因素研究[J]. 西南石油大学学报(社会科学版),2017,19(6):76 –82.
[19] 刘勇,吴翔宇,解本巨. 基于动态用户画像的信息推荐研究[J]. 计算机系统应用,2018,27(6):236 –239.
[20] 王益成,王萍. 基于用户动态画像的科技情报服务推荐模型构建研究[J]. 情报理论与实践,2019,42(4):83 –88.
[21] 张慧敏. 基于生活方式转型的动态用户画像研究[D]. 无锡:江南大学,2018.
[22] 张莉曼,张向先,卢恒,等. 知识直播平台付费用户群体画像研究[J]. 图书情报工作,2019,63(5):84 –91.
[23] 刘漫. 基于 TPB 的大学生信息搜寻行为决定因素实证研究[J]. 图书馆工作与研究,2014(5):39 –44.
[24] LEWIN K. Field theory in social science[M]. New York: Harpperand Brother Publishers, 1951:239 –240.
[25] 许鹏程,毕强,张晗,等. 数据驱动下数字图书馆用户画像模型构建[J]. 图书情报工作,2019,63(3):30 –37.
[26] 衡书鹏,周宗奎,雷玉菊,等. 现实 – 理想自我差异对青少年游戏成瘾的影响:化身认同和沉浸感的序列中介作用[J]. 心理

与行为研究,2018,16(2):253 –260.
[27] 石晓姣. 管理团队认知能力对决策效果的影响机制研究[D]. 沈阳:沈阳工业大学,2018.
[28] 张继东,骆莎莎. 基于情景化偏好的移动图书馆用户行为感知研究[J]. 情报科学,2018,36(9):52 –56.
[29] 杜巍,高长元. 移动电子商务环境下个性化情景推荐模型研究[J]. 情报理论与实践,2017,40(10):56 –61.
[30] 余传明,田鑫,郭亚静,等. 基于行为 – 内容融合模型的用户画像研究[J]. 图书情报工作,2018,62(13):54 –63.
[31] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine learning research, 2003, 3(1): 993 – 1022.
[32] 曾子明,万品玉. 融合演化特征的公共安全事件微博情感分析[J]. 情报科学,2018,36(12):3 –8,51.
[33] 陈添源. 高校移动图书馆用户画像构建实证[J]. 图书情报工作,2018,62(7):38 –46.
[34] 陈娟,吴卓青,邓胜利. 基于层次聚类法的“知乎”用户细分与行为分析[J]. 情报理论与实践,2018,41(7):111 –116.
[35] 林惠惠. 演化聚类算法研究及其应用[D]. 扬州:扬州大学,2017.
[36] CHAKRABARTI D, KUMAR R, TOMKINS A. Evolutionary clustering[C]//Proceedings of the 12th ACM SIGKDD conference on knowledge discovery and data mining. New York:ACM,2006:554 –560.
[37] 王富鹏. 演化聚类研究及其在金融股票市场的应用[D]. 杭州:浙江大学,2014.
[38] 张琳,牟向伟. 基于 Canopy + K-means 的中文文本聚类算法[J]. 图书馆论坛,2018,38(6):113 –119.
[39] 卢建云,朱庆生,吴全旺. 一种启发式确定聚类数方法[J]. 小型微型计算机系统,2018,39(7):1381 –1385
[40] OLIVEIRA M,GAMA J. A framework to monitor clusters evolution applied to economy and finance problems[J]. Intelligent data analysis,2012,16(1):93 –111.
[41] KRITIKOU Y, DEMESTICHAS P, ADAMOPOULOU E, et al. User profile modeling in the context of Web-based learning management systems[J]. Journal of network & computer applications, 2008,31(4):603 –627.

## 作者贡献说明：

张莉曼：论文思路的提出与初稿撰写；
张向先：论文框架的指导与确定；
吴雅威：论文修改与校对；
郭顺利：论文修改与校对。

## Research on the Construction of Dynamic Portrait Model of Social Academic App Users Based on Small Data

Zhang Liman[1] Zhang Xiangxian[1] Wu Yawei[1] Guo Shunli[2]
[1] School of Management, Jilin University, Changchun 130022
[2] Media College, School of Qufu Normal University, Rizhao 276826

Abstract: [**Purpose/significance**] Enrich and expand the theoretical research system of building dynamic portrait of social academic App users based on small data, so as to provide ideas and reference for the social academic App platform to effectively predict the evolution trend of user behavior and improve the precise service level. [**Method/process**] Firstly, based on the deep analysis of concept and characteristics of the small data, combined with the feature of social academic App, this paper from two aspects of user behavior and the surface of deep factors designed dynamic portrait label system. Then collected the small data with strong correlation and high value with the user as the data support of the portrait, and clarified the acquisition and processing method. Finally, it put forward the research method to realize the dynamic portrait and form the overall frame model. [**Result/conclusion**] The construction of dynamic portrait of social academic App users based on small data can effectively refine the granularity of portrait, and improve the lag of previous portrait, which has important reference value for the promotion of accurate service level of social academic App platform under data-driven situation.

Keywords: small data social academic App dynamic portrait behavior prediction